A Multi-model SVR Approach to Estimating the CEFR Proficiency Level of Grammar Item Features

Flanagan, Brendan; Hirokawa, Sachio; Kaneko, Emiko; Izumi, Emi; Ogata, Hiroaki

A Multi-model SVR Approach to Estimating the CEFR Proficiency Level of Grammar Item Features

Brendan Flanagan*, Sachio Hirokawa†, Emiko Kaneko‡, Emi Izumi§ and Hiroaki Ogata*

*Academic Center for Computing and Media Studies
Kyoto University, Japan
Email: {bflanagan.academic,hiroaki.ogata}@gmail.com

†Research Institute for Information Technology
Kyushu University, Japan
Email: hirokawa@.cc.kyushu-u.ac.jp

‡Center for Language Research
Aizu University, Japan
Email: kaneko@u-aizu.ac.jp

§Center for General and Liberal Education
Doshisha University, Japan
Email: eizumi@mail.doshisha.ac.jp

***Abstract*—Analysis of publicly available language learning corpora can be useful for extracting characteristic features of learners from different proficiency levels. This can then be used to support language learning research and the creation of educational resources. In this paper, we classify the words and parts of speech of transcripts from different speaking proficiency levels found in the NICT-JLE corpus. The characteristic features of learners who have the equivalent spoken proficiency of CEFR levels A1 through to B2 were extracted by analyzing the data with the support vector machine method. In particular, we apply feature selection to find a set of characteristic features that achieve optimal classification performance, which can be used to predict spoken learner proficiency.**

I. Introduction

At present there are many machine readable data that are publicly available, and this has increased the application of machine learning to the task of supporting language learning. In this paper, we analyze the NICT-JLE corpus[1] to investigate which words describe and discriminate different speaking proficiency levels by applying a method of machine learning called SVM (Support Vector Machine) to the classification task. The corpus consists of 1280 transcribed recordings of the Standard Speaking Test [1], [2], [3] (herein referred to as SST) English language learner exam. Each exam contains 3 different tasks and the transcriptions are made up of the dialogue between the examiner and examinee. The proficiency level for each examinee was determined by an expert examiner and ranked on a scale from 1 to 9, from beginner to advanced respectively. In this paper, the focus of the classification analysis will be on the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) (Council of Europe, 2001) [4] which is utilized internationally, rather than the SST proficiency levels that are applicable only within Japan. The equivalent proficiency levels of SST, CEFR, and CEFR-J (a version of the CEFR that has been tailored to the needs of Japanese learning English) as defined by Tono et al. [5] are shown in Table I. It should be noted that SST level 4 can be assigned to either CEFR level A1 and A2. The differences in sample sizes across all of the proficiency levels can be see in Table II. In this paper, the evaluation of the classification method was performed with SST level 4 included in the CEFR level A2. The classification of SST level 4 included in the CEFR level A1 should be investigated in future work. SST level 9 is included only in CEFR level B2.

For each of the 1280 examinee's in the SST data there are 5 stages of the interview that have been transcribed. In this paper, the results for each examinee were represented as one document, and there were 1280 sample documents for which the proficiency level classification problem was analyzed. A total of 9,626 words were analyzed along with 11 parts of speech (POS) from Lancaster University's CLAWS5 and CLAWS7 tag sets[2].

Automated language scoring using a computer was first proposed by Page in 1968 [8]. Since then research into the prediction of foreign language proficiency has focused on a

[1]http://alaginrc.nict.go.jp/nict_jle/index_E.html

[2]http://ucrel.lancs.ac.uk/claws5tags.html, http://ucrel.lancs.ac.uk/claws7tags.html

TABLE I
EQUIVALENT LEVELS OF CEFR, CEFR-J, AND SST

CEFR	CEFR-J	SST
-	Pre A1	1
A1	A1.1	2/3
	A1.2	3
	A1.3	4
A2	A2.1	4
	A2.2	5
B1	B1.1	6/7
	B1.2	8
B2	B2.1	9
	B2.2	9
C1	C1	9
C2	C2	9

TABLE II
DATA SAMPLE SIZE

Level	CEFR A1	CEFR A2
A1	738	717
A2	236	257
B1	263	263
B2	40	40

number of different approaches. Supnithi et al. [9], analyzed the vocabulary, grammatical accuracy and fluency features of the NICT-JLE corpus. SVM and Maximum Entropy classifiers were trained to automatically predict the proficiency level of the learner, with SVM achieving the best prediction accuracy of 65.57%. There has also been research into extracting features that can be useful in classifying proficiency levels in the NICT-JLE corpus [10], [11].

Previously we have investigated the same task from the perspective of binary classification. This divided the task into the subtasks of classifying different proficiency levels in the corpus using 1 to 1 class classification. Feature selection was then applied to each of the classifiers to not only improve the performance of the classifier, but also identify a smaller set of characteristic features that accurately describe the classification between a pair of proficiency levels. These features could then be used to assess the proficiency of a document as a binary classification problem, however it can only describe if a feature represent a curtain proficiency level at a local level, and does not provide a global estimation of the difficulty of a feature with respect to proficiency levels. Another method for estimating the difficulty of features with respect to proficiency levels in to train a regression model to predict the proficiency level of a document. However, the proficiency level of a feature is still ambiguous. In this paper, we propose a method for estimating the proficiency level (difficulty) at which a feature exists through the use of staggered Support Vector Regression. The results of our experiment identify both the difficulty and importance of Grammar Item features with respect to the regression of the proficiency level of documents in a corpus of transcribed speaking exams.

TABLE III
5 EXAMPLE GRAMMAR ITEM FEATURES

ID	Grammar Item
1	personal pronoun nominative (I am)
3	personal pronoun nominative (he/she is)
11	instruction adjective (this/that + noun)
137	auxiliary verb (should)
253	wish + subjunctive past

II. PROFICIENCY LEVEL ESTIMATION BY STAGGERED SUPPORT VECTOR REGRESSION

A. Data

The transcripts contained in the NICT-JLE corpus are divided into 5 main stages in the exam. Within stages 2 to 4 there are also tasks and follow-up sections of the stage. The follow-up sections of the exams were excluded from analysis as they contain free dialog between the examiner and examinee. The remaining parts of the corpus were parsed using the method in Tono [7] and Ishii [6] to extract the occurrence of 493 different grammar item features, such as the examples in Table III.

A total of 1280 documents were indexed to form a binary feature vector representation for analysis.

B. Staggered SVR

In order to estimate the difficulty of grammar item features with respect to proficiency level, we use a series of SVR models that are trained at staggered intervals across the proficiency level range from SST level 1 (beginner) to SST level 9 (advanced). Each of the documents in the corpus contain the SST proficiency level of examinee, and we will refer to this as the *original target value*. As the staggered SVR moves across the range of SST levels, this value is altered in relation to the current origin SST level being analyzed. The *target class value* used to train an SVR model at a curtain current origin SST level is calculated by Equation 1.

$$TargetClassValue(d_i, l) = \frac{d_i^{Level} - l}{|L| - 1} \quad (1)$$

Where d_i is the i^{th} document in the corpus, l is the current origin SST level with L representing the set of all SST levels, and d_i^{Level} is the *original target value* of the document d_i. Therefore, when the *original target value* of a document is the same as the current origin SST level, the target class value will be zero. Features that are associated with the current origin SST level will have a strong tendency to have a weight around zero. However this will change as the current origin SST level changes, therefore making it easy to identify features that are associated with a particular level as opposed to a feature that doesn't have discriminative use to the particular regression task.

C. Experiment

A SVR model was trained and evaluated using 10-fold cross validation for each SST level. The prediction performance of each level was measured by: mean average error (MAE), root mean squared error (RMSE), and the accuracy of the model as

TABLE IV
EVALUATION OF SVR MODELS FOR ORIGIN SST LEVELS 1-9

Origin	MAE	RMSE	Accuracy
1	0.0925	0.1254	0.9977
2	0.0930	0.1260	0.9703
3	0.0938	0.1267	0.8234
4	0.0947	0.1278	0.7594
5	0.0958	0.1290	0.8398
6	0.0973	0.1308	0.8773
7	0.0993	0.1332	0.9211
8	0.1011	0.1357	0.9664
9	0.1031	0.1385	1.0000
Average	0.0967	0.1303	0.9061

a binary classifier at the current SST level. The evaluation of the SVR models and the total average across all of the models is shown in Table IV. The change in the accuracy of the models over the SST levels can be associated the differences in the number of samples that are available for each level.

III. ESTIMATION OF GRAMMAR ITEM PROFICIENCY LEVEL AND IMPORTANCE

The feature weights of the SVR model for each origin level were extracted. The weight of the same feature over a series of origin levels can imply the difficulty of the feature by finding when the weight changes polarity. Figure 1 shows the top 10 features whose weight changes from positive to negative as the origin level increases. The point at which a features weight intercepts 0 on the vertical axis represents the proficiency level associated with the feature. The gradient of the weight represents the amount of discriminative use, and therefore importance, that the feature has to the particular regression task. It should also be mentioned that there are also feature weights that change from negative to positive as the origin level increases, as seen in Figure 2 which shows the top 10 positive gradient features.

A. Modeling Grammar Item Feature Weights

To find the gradient, which represents the discriminatory importance of the feature, and intercept point at which a feature weight changes polarity, which represents the proficiency level of the feature, we created least squares regression models for each feature. As we are only modeling the relation between the proficiency level and a features weight, this can be represented by a simple least squares regression model [12] in the form of Equation 2, where b_0 is the bias term, and b_1 is the gradient term. The formula in Equation 3 estimates the gradient term b_1, where $\bar{x}$ and $\bar{y}$ are the mean of all instances of x_i and y_i respectively. Equation 4 estimates the bias term b_0 of the model. The proficiency level of a feature can be estimated by finding the intercept of the regression model of its weights over the range of proficiency levels, as seen in Equation 5. The importance of the feature in discriminating proficiency levels is represented by the rate at which the feature weights change, with greater rate of change indicating that the feature has a strong association with a particular proficiency level. The negative of the gradient term b_1 as shown in Equation 6 can be thought of as the importance of a feature, with larger values representing greater discriminatory importance.

$$y = b_0 + b_1 x \tag{2}$$

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{3}$$

$$b_0 = \bar{y} - b_1 \bar{x} \tag{4}$$

$$ProficiencyLevel(w_i) = \frac{b_0^i}{-b_1^i} \tag{5}$$

$$Importance(w_i) = -b_1^i \tag{6}$$

B. Results

Plots of the top 10 positive and negative slope regression feature weight models are shown in Figures 1 and 2 respectively. Models that have a strong negative gradient are a close fit to the original feature weights on which it was trained. This can also be seen in the evaluation of fit shown in Table V with all of the top 10 models having a R^2 of greater than 0.98. In comparison, the top 10 positive gradient regression models have less of a tight fit to the original feature weights as shown in Table VI with all top 10 models having a R^2 of only greater than 0.86. It should be noted that a majority of the positive gradient models are associated with proficiency levels that are outside the normal SST and CEFR-J scales.

Several features listed in Tables V and VI are elements that realize advanced utterance and relate to the following: complexity of the utterance (155: adverbial clause "as soon as", 175: complex relative pronoun "what", 166: present participle for post-qualifying nouns, 189: S + V "give/pass/send/show/teach/tell" + IO + DO), distinct functions in communication (241: function Question "Can you ...?", 139-2: auxiliary verbs "would", 254: function question "How about ...?"), expressing subtle nuances (139-2: auxiliary verbs "would"), and indication of relationship with other elements in utterance (17: determinant "another").

An overview of all of the feature weight model analysis is shown in Figure 3, with the level of Grammar Items represented on the x-axis, and the y-asix represents the discriminatory importance of the feature. It should be noted that some features are not shown in the plot because the level of the feature was far from the normal level range. A majority of the features that have a relatively high level of importance are within the upper beginner to intermediate level range. There are also numerous features with relatively low importance in the lower levels.

IV. CONCLUSION

Previous research into estimating features that can discriminate between different proficiency levels have provided positive or negative feature sets with respect to the classification problem. In this paper, we propose a method of estimating the level of a feature in respect to whole proficiency ranges by applying staggered SVR, which provides a tangible level as opposed to previous work. Our method can also identify the discriminatory importance of a feature, which could be used

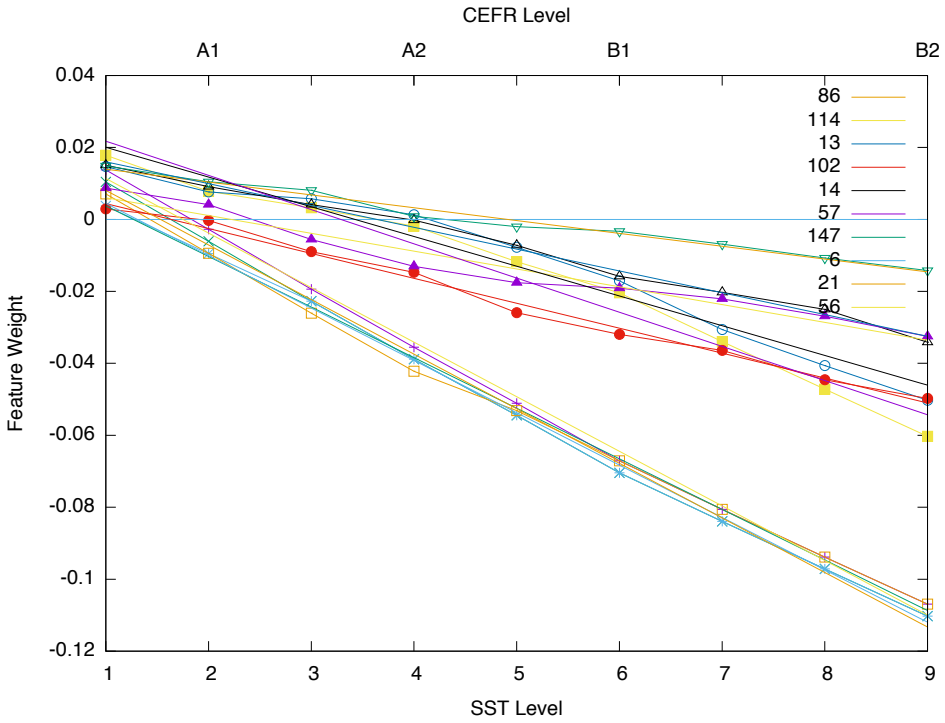

Fig. 1. Top 10 negative gradient feature weights and related least squares regression plots.

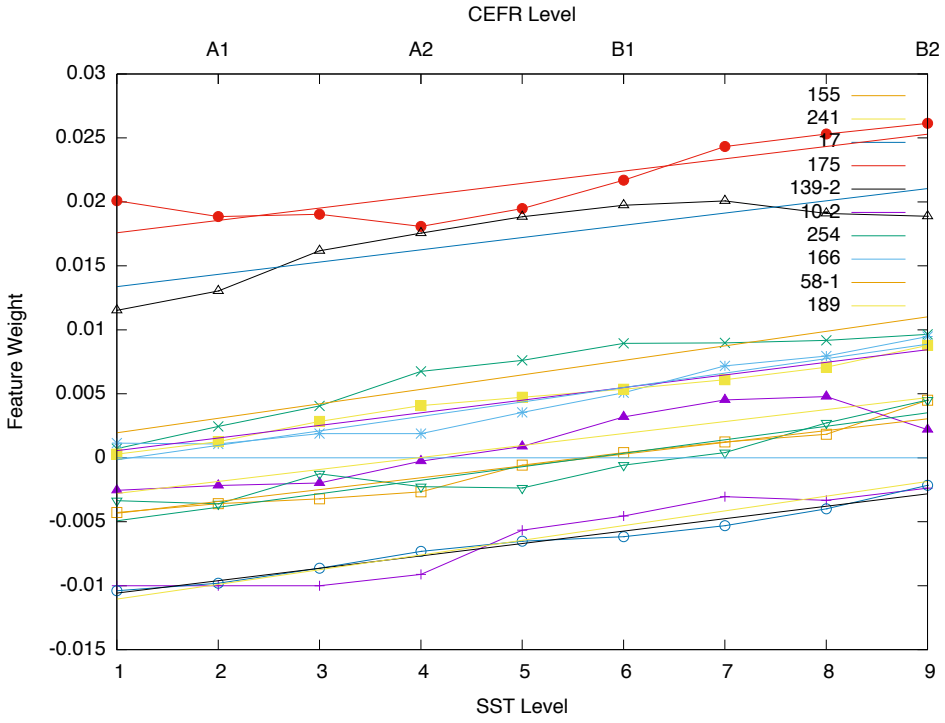

Fig. 2. Top 10 positive gradient feature weights and related least squares regression plots.

TABLE V
TOP 10 NEGATIVE SLOPE REGRESSION MODELS.

#:Grammar Item	SST	CEFR	b_0	b_1	R^2
56:TA.PRESENT.be.AFF	1.7509	A1	0.0265	-0.0151	-0.9989
21:II.PREP.GENERAL	1.5311	A1	0.0232	-0.0151	-0.9989
6:APPGE	1.3069	A1	0.0190	-0.0145	-0.9994
147:CC	1.2670	A1	0.0178	-0.0140	-0.9985
57:TA.PRESENT.do.AFF	3.2872	A1	0.0312	-0.0095	-0.9863
14:AT.the	3.4258	A1	0.0283	-0.0082	-0.9811
102:VVG	1.6349	A1	0.0113	-0.0069	-0.9961
13:AT.a.an	3.6383	A1	0.0220	-0.0060	-0.9969
114:IMP.VV.AFF	2.2179	A1	0.0109	-0.0049	-0.9807
86:TO.to_do	4.9115	A2	0.0175	-0.0035	-0.9923

TABLE VI
TOP 10 POSITIVE SLOPE REGRESSION MODELS.

#:Grammar Item	SST	CEFR	b_0	b_1	R^2
189:VP.SVOO.AFF	>9	>B2	-0.0122	0.0012	0.9543
58-1:TA.PRESENT.does.NEG	<1	<A1	0.0008	0.0011	0.9438
166:VVG.N_VVG	1.1377	A1	-0.0013	0.0011	0.9681
254:INTF.how_about	5.6689	A2	-0.0060	0.0010	0.9791
10-2:PPH1.is_it	<1	<A1	-0.0004	0.0010	0.9913
139-2:VM.would.INT.AFF	>9	>B2	-0.0115	0.0010	0.9894
175:PNQ.REL.what	<1	<A1	0.0166	0.0010	0.8621
17:DD1.another	<1	<A1	0.0124	0.0009	0.8607
241:INTF.can_you	3.9794	A1	-0.0037	0.0009	0.8945
155:CS.as_soon_as	5.6963	A2	-0.0053	0.0009	0.9158

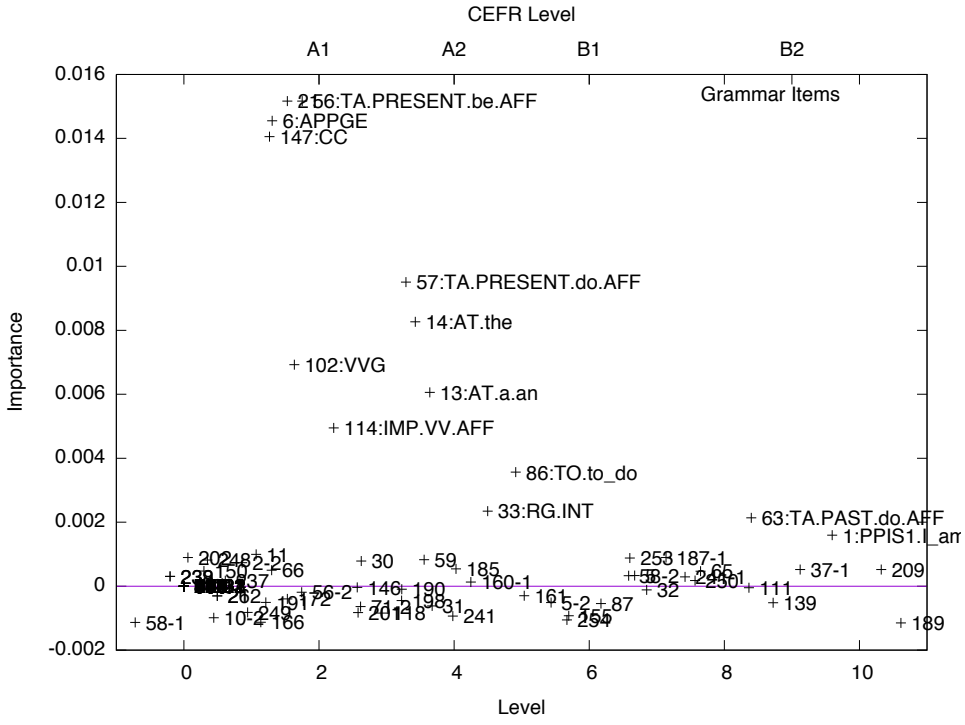

Fig. 3. Grammar item feature level (with CEFR equivalent scale) versus the importance of each feature.

to rank features within a level. In future work, we plan to investigate the methods for improving the performance of SVR by applying feature selection, and also identify an optimal subset of features that represents the whole proficiency level range effectively.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 15J04830 and 16H06304.

REFERENCES

[1] Izumi, E., Uchimoto, K., Isahara, H.: The NICT JLE corpus, ACL Publishing, 2004. (in Japanese)
[2] Izumi, E., Uchimoto, K., Isahara, H.: The NICT JLE Corpus: Exploiting the language learner's speech database for research and education. International Journal of the Computer, the Internet and Management 12(2), 119–125, 2004.
[3] Izumi, E., Uchimoto, K., Isahara, H.: The Overview of the SST Speech Corpus of Japanese Learner English and Evaluation through the Experiment on Automatic Detection of Learners' Errors, LREC, 1435–1438, 2004
[4] Council of Europe: Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press, 2001
[5] Yukio Tono (Ed.): The CEFR-J handbook: a resource book for using CAN-DO descriptors for English language teaching, Taishukan Publishing (2013) (in Japanese)
[6] Ishii, Y., Tono, Y., Grammatical Items for Creating the CEFR-J Grammar Profile, A Symposium on the CEFR-J RLD Project: Developing Grammar, Text and Error Profiles Using Textbook & Learner Corpora, Tokyo Foreign Studies University, 2016.
[7] Tono, Y., Ishii, Y., How to use CEFR-J Grammar Profile, A Symposium on the CEFR-J RLD Project: Developing Grammar, Text and Error Profiles Using Textbook & Learner Corpora, Tokyo Foreign Studies University, 2016.
[8] Page, E.B., The use of the computer in analyzing student essays, International Review of Education 14(2), pp. 210–225, 1968.
[9] Supnithi, T., Uchimoto, K., Saiga, T., Izumi, E., Virach, S., Isahara, H., Automatic proficiency level checking based on SST corpus, Proc. RANLP, pp. 29–33, 2003.
[10] Abe, M., Frequency Change Patterns across Proficiency Levels in Japanese EFL Learner Speech, Apples: Journal of Applied Language Studies 8(3), pp. 85–96, 2014.
[11] Flanagan, B., Hirokawa, S., The Relationship of English Foreign Language Learner Proficiency and an Entropy Based Measure, IEE 1(3), pp. 29–38, 2015.
[12] Rao, C.R., Toutenburg, H., Shalabh, Heumann, C.: Linear models and generalizations: Least Squares and Alternatives (3rd ed), Springer, 2008